The interplay of mutations and electronic properties in disease-related genes

Chi-Tin Shih[1], Stephen A. Wells[2], Ching-Ling Hsu[3] Yun-Yin Cheng[1], & Rudolf A. Römer[2,*]

[1] *Department of Physics, Tunghai University, 40704 Taichung, Taiwan and The National Center for Theoretical Sciences, 30013 Hsinchu, Taiwan*

[2] *Department of Physics and Centre for Scientific Computing, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK*

[3] *Department of Physics, Chung-Yuan Christian University, 32023 Chung-Li, Taiwan*

[*] *Correspondence: r.roemer@warwick.ac.uk*

Electronic properties of DNA are believed to play a crucial role in many phenomena in living organisms, for example the location of DNA lesions by base excision repair (BER) glycosylases and the regulation of tumor-suppressor genes such as p53 by detection of oxidative damage. However, the reproducible measurement and modelling of charge migration through DNA molecules at the nanometer scale remains a challenging and controversial subject even after more than a decade of intense efforts. Here we show, by analysing 162 disease-related genes from a variety of medical databases with a total of almost $20,000$ observed pathogenic mutations, a significant difference in the electronic properties of the population of observed mutations compared to the set of all possible mutations. Our results have implications for the role of the electronic properties of DNA in cellular processes, and hint at the possibility of prediction, early diagnosis and detection of mutation hotspots.

Cells tend to accumulate over time genetic changes such as nucleotide substitutions, small insertions and deletions, rearrangements of the genetic sequences and copy number changes.[1] These changes in turn affect protein-coding or regulatory components and lead to health issues such as cancer, immunodeficiency, ageing-related diseases and other disorders. A cell responds to genetic damage by initiating a repair process or programmed cell death.[2] In recent years, a vast number of detailed databases have been assembled in which rich information about the type, severity, frequency and diagnosis of many thousand of such observed mutations has been stored.[3–6] This abundance of data is based on the now standard availability of massively parallel sequencing technologies.[7] Harvesting these genomic databases for new cancer genes and hence potential therapeutic targets has already demonstrated its usefulness[8] and several recent international cancer genome projects continue the required large-scale analysis of genes in tumours.[9]

The possible relevance of charge transport in DNA damage has recently also attracted considerable interest in the bio-chemical and bio-physical literature.[10–13] Direct measurement of charge transport and/or transfer in DNA remains a highly controversial topic due to the very challenging level of required manipulation at the nano-scale.[14] Ab-initio modelling of long DNA strands is similarly demanding of computational resources and so some of the most promising computational approaches necessarily use much simplified models based on coarse-grained DNA.[11] Here we compute and datamine the results of charge transport calculations based on two such effective models for each possible mutation in 162 of the most important disease-associated genes from four large gene databases. The models are (i) the standard one-dimensional chain of coupled nucleic bases with onsite ionisation potentials[11,15] as well as a novel 2-leg ladder model with diagonal

couplings and explicit modelling of the sugar-phosphate backbone.[16]

Results

Point Mutations and Electronic Properties We consider native genetic sequences and mutations of disease-associated genes as retrieved from the *Online Mendelian Inheritance in Man* (OMIM)[3] of NCBI, the *Human Gene Mutation Database* (HGMD),[4] *International Agency of Research on Cancer* (IARC)[5] as well as *Retinoblastoma Genetics*.[6] We have selected these genes such that (i) those from OMIM have a well-known sequence with known phenotype as well as at least 10 point mutations, (ii) all other selected cancer-related genes have also at least 10 point mutations and (iii) all non-cancer related genes from HGMD have at least 200 point mutations (cp. Supplementary Table S1).

Many different types of mutation are possible in a genetic sequence including point mutations, deletion of single base pairs (producing a frame shift), and large-scale deletion or duplication of multiple base pairs. Here, we restrict our attention to point mutations as it allows us to directly compare the sequence before and after the mutation. We study the magnitude of the *change* in charge transport (CT) for pathogenic mutations when compared to all possible mutations either *locally*, i.e. at the given hotspot site, or *globally* when ranked according to magnitude of CT change. We find that the vast majority of mutations shows good agreement with a hypothesis where *smallest change in electronic properties* — as measured by a change in CT — *corresponds to a mutation* that has appeared in one of the aforementioned databases *of pathogenic genes*.

A gene with $\mathcal{N}$ base pairs (bps) has a native nucleotide sequence $(s_1, s_2, \cdots, s_{\mathcal{N}})$ along the coding strand. The gene has a total of $3\mathcal{N}$ possible point mutations, which we denote as the set $M_{\rm all}$, of which a subset $M_{\rm pa}$ are known pathogenic mutations. A point mutation is represented by the pair (k, s), where k is the position of the point mutation in the genomic sequence and s is the mutant nucleotide which replaces the native nucleotide. We shall write a mutation from a native base P to a mutant base Q as "Pq". We note that there are a total of twelve possible point mutations in a DNA sequence (from any one of four bases to any one of three alternatives). Of these twelve, four are *transitions*, in which a purine base replaces a purine or a pyrimidine replaces a pyrimidine, and eight are *transversions* in which purine is replaced by pyrimidine or vice versa. Biologically, transitions are in general much more common than transversions.[17] Indeed, the set of observed pathogenic mutations for our 162 genes contains 10999 transitions and 8883 transversions, whereas in the set of all mutations their ratio is by definition $1:2$. The observed pathogenic mutations are thus already a biased selection from the set of possible mutations, favouring transitions. However, this local onsite chemical shift is not sufficient to fully explain our data as we will show later.

We compute and datamine the results of quantum mechanical transport calculations based on two effective Hückel models[18] for each possible mutation in those 162 genes. Both models assume π–π orbital overlap in a well-stacked double helix. The parameters are chosen to represent hole transport. Using the transfer matrix method[19,20] we calculate the spatial extent of (hole) wavefunctions of a given energy on a length of DNA with a given genetic sequence. Wavefunction localisation is directly related to conductance[19] and we therefore find it convenient to report our results in terms of conductance.

To determine the effect of a mutation, we consider sub-sequences of length L bps; there are L such sequences that include a given site k. For all L sequences we calculate quantum-mechanical charge transmission coefficients T (in units of $e^2/\hbar$, averaged across a range of incident energies, as detailed in Methods) for the native and mutant sequences. We describe the effect of the mutation on the electronic properties of the DNA strand near to the mutation site using the mean square difference, $\Gamma = \langle |T_{\rm native} - T_{\rm mutant}|^2 \rangle$, averaged across all L sequences. Larger values of Γ therefore correspond to a greater difference in electronic structure between the native and mutant sequences. The length L must be long enough to allow for substantial delocalisation across multiple base pairs,[21] but should remain below the typical persistence length of ~ 150 bps[22] such that any overlap or crossing by packing, e.g. by wrapping around histone complexes in chromatin, can be ignored. In this study we have considered lengths of $20, 40, 60$ bps. This requires, for each of the $\mathcal{N}$ sites in a gene, L calculations for each sequence of length L and for each of 4 possible bases at that site; which, for the more than 11×10^6 bases in our dataset of 162 genes, is more than 5×10^9 quantum mechanical transport calculations.

Local and global ranking We first compare Γ of each observed pathogenic mutation with the other two non-pathogenic ones at the same position and determine a *local ranking* (LR) of CT change. There are three possibilities of LR, namely *low*, *medium* and *high*. Note that those hotspots with more than one pathogenic mutations are excluded in the LR analysis. We have also sorted the LR ranking for each gene according to prevalence in Fig. 1(a+b). We find that for $L = 20, 40$ and 60 the low CT change corresponds to 155 (95%), 148 (91%) and 140 (86%) of all 162 genes with pathogenic mutations. Examples of LR for the pathogenic mutations of *p16* and *CYP21A2*

are shown in Supplementary Fig. S3. We graphically summarise the results for all 162 disease-associated genes in Fig. S5. For each gene, we have shown a positive deviation from the 33% line by orange —supporting the scenario of small CT change for pathogenic mutations — and by blue when the results seem to show no or negative indication with CT change. It is clear that the correlation between low CT change and mutation hotspots is well pronounced.

We can also consider a *global* ranking (GR) by sorting CT change Γ for *all* possible $3\mathcal{N}$ mutations of a gene with $\mathcal{N}$ bps in order to get a ranking of *every* observed pathogenic mutation. By dividing each ranking by $3\mathcal{N}$ we compute the normalised GR γ of the mutation, with values between 0 and 1. Smaller values of γ mean smaller CT change. By analogy to the local ranking, we divide the γ of the pathogenic mutations into three groups as before, i.e. low ($\gamma < 33.3\%$), medium ($33.3\% \leq \gamma < 66.7\%$), and high ($\gamma \geq 66.7\%$) CT change. The results of the GR for the 162 genes are shown in the bottom row (c) and (d) of Fig. 1. As for the LR results, we observe many γ values with low CT change (cp. Supplementary Figs. S3and S4). Hence the LR and GR results consistently show that observed pathogenic mutations are generally biased towards smaller change in CT than the set of all possible mutations (cp. Supplementary Figs. S5and S6).

Distributions of change in charge transport In Figure 2 we show as an example results for the distribution of Γ for the $p16$ DNA strand for both 1D and 2-leg models. In panels (a+b), it is clear that the 111 observed pathogenic mutations of $p16$ have on average *smaller changes* in the CT properties as compared to all possible 80220 mutations, for both the 1D and 2-leg models. We find that results for the vast majority of the other 161 genes are quite similar. The distributions of Γ

values in Fig. 2(a+b) are approximately log-normal. We therefore calculate, for each of the 162 genes in our dataset, an average $\log\Gamma$ value for the distributions of all and pathogenic mutations. Histograms of the distributions of these $\langle\log\Gamma\rangle$ values are shown in Fig. 2(c+d). It is once again clear that the distributions for observed pathogenic mutations are shifted towards lower Γ values in both the 1D and the 2-leg models.

We next define a *global CT shift* for a gene g as $\Lambda_g = \langle\log\Gamma_{g,\mathrm{all}}\rangle - \langle\log\Gamma_{g,\mathrm{pa}}\rangle$. Positive values of Λ_g indicate that the observed pathogenic mutations of gene g have a lower average Γ. For each of our 162 genes we obtain the distribution of Λ_g for the 1D and 2-leg models as shown in Figs. 2(e+f). We can define, for the whole set of 162 genes, an average global shift $\bar{\Lambda} = \sum_g \Lambda_g/162$, weighting all genes equally; we can also weight the results by the number of observed pathogenic mutations for each gene $|M_{\mathrm{pa}}|_g$ for a *weighted* average global shift $\tilde{\Lambda} = \frac{1}{\sum_g |M_{\mathrm{pa}}|_g}\sum_g |M_{\mathrm{pa}}|_g\Lambda_g$. These values are also indicated in Figs. 2(e+f) and in both models there is a tendency towards lower average $\bar{\Lambda}_g$ for observed pathogenic mutations.

Transitions and transversions In our models we would expect transitions to cause, in general, a smaller change in CT than transversions, as the change in onsite energy and in transfer coefficients is smaller for a transition than a transversion. However, as we will demonstrate here, the increased proportion of transitions among the observed pathogenic mutations is *not* sufficient to account for the distributions seen in Fig. 2.

In Fig. 3(a+b) we show the distribution of Γ values for our entire dataset of all $\simeq 34\times10^6$ possible mutations and 19882 known pathogenic mutations, dividing the datasets into transitions

and transversions. For both models, the transitions are shifted to slightly lower Γ values than the transversions. However, in the 2-leg model, the distribution for observed pathogenic transitions appears co-located with the distribution for all transitions, and likewise for transversions. In the 1D model, by contrast, the observed pathogenic transitions are visibly shifted to lower Γ values than the set of all transitions, and the same is true for transversions.

In Fig. 3(c+d) we represent the distributions of Γ values for each of the twelve types of point mutation by points for the mean values of $\log \Gamma$ and bars indicating the standard deviation of the distribution of $\log \Gamma$. In the 2-leg model, the distributions for observed pathogenic mutations are essentially coincident with the distributions for all mutations for each type Pq. The positive $\bar{\Lambda}$ and $\tilde{\Lambda}$ shift results in the 2-leg model are thus accounted for by the set of observed pathogenic mutations being biased towards transitions. The 1D model displays a quite different behaviour; in each case the mean of the distribution for the observed pathogenic mutations of any type Pq, lies from 7.5 to 20 standard errors *below* the mean for all possible mutations of type Pq. Hence the probability that the observed pathogenic mutations are a random subset of all mutations, with respect to their electronic properties in the 1D model, is comparable to the probability of drawing twelve values more than 7.5 standard deviations below the mean from a normal distribution, which is less than 10^{-168}. The observed difference between CT change between observed pathogenic and all possible mutations is thus statistically highly significant irrespective of whether transitions or transversions are involved. In the 2D model, by contrast, the means of the $\log \Gamma$ distributions for observed pathogenic mutations can lie either above or below those for all mutations for different types Pq, and the difference in the means — between 0.03 and 5.5 standard errors — is much

smaller.

Let us also consider, for each gene g, simulation length L and each mutation type Pq whether the *subset shift* $\lambda = \langle \log \Gamma_{\rm all} \rangle - \langle \log \Gamma_{\rm pa} \rangle_{g,L,{\rm Pq}}$ is positive or negative. This gives us, for each model, $162 \times 3 \times 12 = 5832$ data points, less 1029 cases where no calculation is possible as no pathogenic mutations of type Pq are known for gene g. These λ data are presented in Fig. 4. In the 2-leg model there are approximately equal numbers of negative and positive λ values. This is consistent with a null hypothesis where the observed pathogenic mutations of a type Pq have the same distribution of Γ vales as for all mutations of that type. In the 1D model, by contrast, such a null hypothesis is decisively rejected: there is a preponderance of positive λ values by almost $2:1$ (3326 positive to 1513 negative) and the binomial probability of obtaining such a result at random would be approximately 10^{-153}. The two analyses agree that observed pathogenic mutations display a significant bias towards smaller changes in electronic properties in the 1D model.

Discussion

Our CT models act as probes of the statistics of the DNA sequence. It is possible that we are merely observing a correlation; i.e. that mutations are more likely to occur in areas of the genome with certain statistical properties, for reasons not causally related to charge transport, and these properties correlate with biased CT properties in our 1D model. Such a correlation between quantum transport and mutation hotspots would in itself be a valuable and novel observation in bioinfor-

matics. There are known chemical biases in the occurence of mutations, such as the enhanced transition rate in C-G doublets,[23] the bias towards GC base pairs rather than AT pairs in biased gene conversion[24,25] and the tendency of holes to localise on GG and GGG sequences and there cause oxidative damage.[26] However, since our observed bias is consistent across all twelve types of point mutation, these known biases cannot fully account for our data.

There are also plausible causal connections between our data and cellular genetic processes where the electronic properties of DNA may be significant. One such process is gene regulation, where charge transport along the DNA strand can couple to redox processes in DNA-bound proteins, inducing protein conformational change and unbinding.[27] Similarly, it has been proposed that DNA repair glycosylases containing redox-active [4Fe-4S] clusters[28] may localise to the site of DNA lesions through a DNA-mediated charge transport mechanism.[29] The recognition of specific areas in the DNA sequence by DNA-binding proteins generally may involve electrostatic recognition of the target DNA sequence.[30] Furthermore, homologous recombination[31] — a process which is vital to the repair of double-strand breaks, a most serious DNA lesion,[32,33] and also to genetic recombination — relies on the mutual recognition of homologous chromosomes before strand invasion can occur. Homologous double-stranded DNA sequences are capable of mutual recognition even in a protein-free environment,[34] presumably via electronic or electrostatic interactions.[35]

All the above processes, especially those involving protein–DNA or DNA–DNA recognition, would be less disrupted by a smaller change in the electronic environment along the coding strand. From this point of view, the observed mutations are biased to cause *less* disruption to gene regula-

tion and DNA damage repair in the cell. This may seem counterintuitive at first. However, in order for a mutation to appear in our dataset of pathogenic mutations, the cell and the organism must develop viably for long enough for a mutant phenotype to be observed. Mutations which cause large disruptions to DNA regulation and repair are more likely to be lethal to the cell at an early stage and will thus be absent from disease databases. Similarly, mutations which are more visible to DNA repair mechanisms are less likely to persist and to appear in databases.

Genetic repair and regulation mechanisms cannot know whether the consequences of a mutation are beneficial, neutral or harmful. We would therefore predict that neutral mutations should display the same bias, towards smaller change in electronic structure, as we observe in the pathogenic mutations. As a first test of this prediction, we have considered the case of the TP53 gene, with 20303 base pairs and for which there are known 2003 pathogenic mutations, 366 silent mutations and 113 intronic mutations.[5] We have simulated these silent and intronic mutations using the 1D model. Histograms of the distribution of Γ values for these mutations are given in supplementary material, see Fig. S7. In Table 1 we analyze the statistical properties for the resulting Γ distributions; our results demonstrate that, for both transitions and transversions, the silent and intronic mutations are similar to the pathogenic mutations and significantly *disimilar* to the population of all possible mutations, as predicted.

Methods

Models of charge transport in DNA. The simplest model of coherent hole transport in DNA is given by an effective one-dimensional Hückel-Hamiltonian for CT through nucleotide HOMO states,[11] where each lattice point represents a nucleotide base (A,T,C,G) of the chain for $n = 1, \ldots, N$. In this tight-binding formalism, the on-site potentials ϵ_n are given by the ionisation potentials $\epsilon_{\rm G} = 7.75e$V, $\epsilon_{\rm C} = 8.87e$V, $\epsilon_{\rm A} = 8.24e$V and $\epsilon_{\rm T} = 9.14e$V, at the nth site, cp. Fig. 5; the hopping integrals $t_{n,n+1}$ are assumed to be nucleotide-independent with $t_{n,n+1} = 0.4e$V.[11] A model which is less coarse-grained is provided by the diagonal, 2-leg ladder model shown in Fig. 5. Both strands of DNA and the backbone are modelled explicitly and the different diagonal overlaps of the larger purines (A,G) and the smaller pyrimidines (C,T) are taken into account by suitable inter-strand couplings.[16,36] The intra-strand couplings are $0.35e$V between identical bases and $0.17e$V between different bases; the diagonal inter-strand couplings are $0.1e$V for purine-purine, $0.01e$V for purine-pyrimidine and $0.001e$V for pyrimidine-pyrimidine. Perpendicular couplings to the backbone sites are $0.7e$V, and perpendicular hopping across the hydrogen bond in a base pair is reduced to $0.005e$V.

The 2-leg model[16] allows inter-strand coupling between the purine bases in successive base pairs, in accordance with electronic structure calculations [36], and should therefore be a better model for bulk charge transport along the DNA double helix; the 1D model, by contrast, makes use of the site energies of only the bases on the coding strand, [15] and so is most representative of the electronic environment along that strand. We also find that the 2-leg model recovers some of the

coding strand dependence of the 1D model upon decreasing the diagonal hoppings. For 28 genes, we find that reducing only the diagonal hopping elements by two leads to a much greater agreement with the 1D results similar to Fig. 3(c).

Calculation of quantum transmission coefficients. The quantum transmission coefficient $T(E)$ for a DNA sequence with length N bps for different injection energy E can be calculated for both models by using the transfer matrix method.[20,37] Let us define $T_{j,L}(E)$ as the transmission coefficient for a part of a given DNA sequence which starts at base pair position j and is L base pairs long. The *position-dependent averaged transmission coefficient* at the $k-$th base pair for transmission length L bps is defined as

$$T_L^{(k)} = \frac{1}{L} \sum_{j=k-L+1}^{k} \int_{E_0}^{E_1} \frac{T_{j,L}(E)}{E_1 - E_0} dE \quad . \tag{1}$$

Here j ranges from $k - L + 1$ to k such that each subsequence of length L contains the kth base pair. E_0 and E_1 are the lower and upper bounds of the incident energy of the carriers, e.g. for the $1D$ model used here, the values are 5.75 and $9.75e$V, respectively; for the 2-leg model the bounds are 7 and $11e$V. We have used an energy resolution of $\Delta E = 0.005e$V. Then we examine the difference between transmission coefficients of the normal and mutated genomic sequence of a point mutation[15] and hence denote by $T_{j,L}^{(k,s)}$ the transmission coefficient of the same segment of DNA as $T_{j,L}^{(k)}$ but with the point mutation (k, s). $\Gamma_L^{(k,s)}$ is the averaged effect of the point mutation (k, s) on CT properties for all subsequences of length L containing the mutation,

$$\Gamma_L^{(k,s)} = \frac{1}{L} \sum_{j=k-L+1}^{k} \int_{E_0}^{E_1} \frac{|T_{j,L}(E) - T_{j,L}^{(k,s)}(E)|^2}{E_1 - E_0} dE \quad . \tag{2}$$

1. Sherbet, G. V. *Genetic Recombination in Cancer* (Academic Press, 2003).

2. Frank, S. A. *Dynamics of Cancer: Incidence, Inheritance and Evolution*. Princeton Series in Evolutionary Biology (Princeton University Press, Princeton and Oxford, 2007).

3. McKusick-Nathans Institute of Genetic Medicine. Online Mendelian inheritance in man (2010). URL `http://www.ncbi.nlm.nih.gov/omim/`. Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

4. Steson, P. D. *et al.* Human gene mutation database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003). URL `http://www.hgmd.cf.ac.uk/ac/index.php`.

5. Petitjean, A. *et al.* Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum. Mutat.* **28**, 622–29 (2007). Http://www-p53.iarc.fr/index.html, R11.

6. Lohmann, D. R. & Gallie, B. A. L. Retinoblastoma: Revisiting the model prototype of inherited cancer. *Am. J. Med. Genet. C* **129C**, 23–28 (2005). Http://www.verandi.de/joomla.

7. Nagl, S. (ed.) *Cancer Bioinformatics* (Wiley, Chichester, England, 2006).

8. Enkemann, S. A., McLoughlin, J. M., Jensen, E. H. & Yeatman, T. J. Whole-genome analysis of cancer. In Gordon, G. J. (ed.) *Cancer Drug Discovery and Development*, chap. 3, 25–55 (Humana Press, 2009).

9. The International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).

10. Starikov, E. B., Lewis, J. P. & Tanaka, S. (eds.) *Modern Methods for Theoretical Physical Chemistry of Biopolymers* (Elsevier, Amsterdam, 2006).

11. Chakraborty, T. (ed.) *Charge Migration in DNA: Perspectives from Physics, Chemistry and Biology* (Springer Verlag, Berlin, 2007).

12. Berashevich, J. & Chakraborty, T. Mutational hot spots in DNA: where biology meets physics. *Physics in Canada* **63**, 103–107 (2007).

13. Genereux, J., Boal, A. & Barton, J. DNA-mediated charge transport in redox sensing and signalling. *J. Am. Chem. Soc.* **132**, 891–905 (2010).

14. Guo, X., Gorodetsky, A. A., Hone, J., Barton, J. K. & Nuckolls, C. Conductivity of a single DNA duplex bridging a carbon nanotube gap. *Nature Nanotechnology* **3**, 163 (2008).

15. Shih, C.-T., Roche, S. & Römer, R. A. Point-mutation effects on charge-transport properties of the tumor-suppressor gene p53. *Phys. Rev. Lett.* **100**, 018105 (2008).

16. Wells, S. A., Shih, C.-T. & Römer, R. A. Modelling charge transport in DNA using transfer matrices with diagonal terms. *Int. J. Mod. Phys. B* 4138–4149 (2009).

17. Collins, D. & Jukes, T. Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* **20**, 386–396 (1994).

18. Powell, B. J. An introduction to effective low-energy hamiltonians in condensed matter physics and chemistry (2010). ArXiv:0906.1640v7 [physics.chem-ph].

19. Kramer, B. & MacKinnon, A. Localization: theory and experiment. *Rep. Prog. Phys.* **56**, 1469–1564 (1993).

20. Ndawana, M. L., Römer, R. A. & Schreiber, M. Effects of scale-free disorder on the Anderson metal-insulator transition. *Europhys. Lett.* **68**, 678–684 (2004).

21. Klotsa, D. K., Römer, R. A. & Turner, M. S. Electronic transport in DNA. *Biophys. J.* **89**, 2187–2198 (2005).

22. Hegerman, P. J. Flexibility of DNA. *Ann. Rev. Biophys. Biophys. Chem* **17**, 265–286 (1988).

23. Blake, R., Hess, S. & Nicholson-Tuell, J. The influence of nearesst neighbors on the rate and pattern of spontaneous point mutations. *J Mol Evol* **34**, 189–200 (1992).

24. Galtier, N. & Duret, L. Adaptation of biased gene conversion? extending the null hypothesis of molecular evolution. *TRENDS in Genetics* **23**, 273–277 (2007).

25. Marais, G. Biased gene conversion; implications for genome and sex evolution. *TRENDS in Genetics* **19**, 330–338 (2003).

26. Nunez, M., Holmquist, G. & Barton, J. Evidence for DNA charge transport in the nucleus. *Biochemistry* **40**, 12465–12471 (2001).

27. Augustyn, K. E., Merino, E. J. & Barton, J. K. A role for DNA-mediated charge transport in regulating p53: Oxidation of the DNA-bound protein from a distance. *Proc. Nat. Acad. Sci.* **104**, 18907–18912 (2007). URL

http://www.pnas.org/cgi/content/abstract/0709326104v1.

http://www.pnas.org/cgi/reprint/0709326104v1.pdf.

28. Boal, A., Yavin, E. & Barton, J. DNA repair glycosylases with a [4fe-4s] cluster: a redox cofactor for DNA-mediated charge transport? *J. Inorg. Biochem.* **101**, 1913–1921 (2007).

29. Yavin, E., Stemp, E. D. A., O'Shea, V. L., David, S. S. & Barton, J. K. Electron trap for DNA-bound repair enzymes: A strategy for DNA-mediated signaling. *Proc. Nat. Acad. Sci.* **103**, 3610 (2006).

30. Cherstvy, A., Kolomeisky, A. & Kornyshev, A. Protein-DNA interactions; reaching and recognizing the targets. *J. Phys. Chem. B* **112**, 4741–4750 (2008).

31. Ferguson, D. & Alt, F. DNA double strand break repair and chromosomal translocation: lessons from animal models. *Oncogene* **20**, 5572–5579 (2001).

32. Jackson, S. Sensing and repairing DNA double-strand breaks- commentary. *Carcinogenesis* **23**, 687–696 (2002).

33. Khanna, K. & Jackson, S. DNA double-strand breaks: signalling, repair and the cancer connection. *Nature Genetics* **27**, 247–254 (2001).

34. Baldwin, G. S. *et al.* DNA double helices recognize mutual sequence homology in a protein free environment. *J. Phys. Chem. B* **114**, 1060–1064 (2008).

35. Kornyshev, A. A. & Leikin, S. Sequence recognition in the pairing of DNA duplexes. *Phys. Rev. Lett.* **86**, 3666–3669 (2001).

36. Rak, J., Voityuk, A., Marquez, A. & Rösch, N. The effect of pyrimidine bases on the hole-transfer coupling in DNA. *J. Phys. Chem. B* **106**, 7919–7926 (2002).

37. Roche, S. Sequence dependent DNA-mediated conduction. *Phys. Rev. Lett.* **91**, 108101–4 (2003).

38. Shih, C. T. Characteristic length scale of electric transport properties of genomes. *Phys. Rev. E* **74**, 010903(R) (2006). URL `http://link.aps.org/abstract/PRE/v74/e010903`.

Supplementary Information available

Acknowledgements This work was supported by the National Science Council in Taiwan (CTS, Grant No. 97-2112-M-029-002-MY3) and the UK Leverhulme Trust (RAR, SAW, Grant No. F/00215/AH). Part of the calculations were performed at the National Center for High-Performance Computing in Taiwan. We are grateful for their help.

Author Contributions CTS and RAR coordinated the international collaboration and wrote the main manuscript text. CTS, RAR and SAW wrote the programs and performed the main computation. YYC and CLH analyzed the source databases and performed the data preprocessing. All authors analyzed the data and reviewed the manuscript.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to CTS (email: ct-shih@thu.edu.tw) or RAR (email: r.roemer@warwick.ac.uk).

	$\overline{\log_{10}\Gamma}$	SEM	σ	$p_{\rm all}$	$p_{\rm pa}$
All transitions	-1.753	0.003	0.427	-	-
Pathological transitions	-1.840	0.015	0.431	$1.01\times^{-8}$	-
Silent transitions	-1.868	0.029	0.440	6.62×10^{-5}	0.391
Intron transitions	-1.805	0.048	0.391	0.320	0.526
All transversions	-1.605	0.002	0.422	-	-
Pathological transversions	-1.710	0.012	0.4190	$< 10^{-10}$	-
Silent transversions	-1.691	0.036	0.432	0.016	0.610
Intron transversions	-1.739	0.054	0.337	0.032	0.636

Table 1: Mean logarithm of CT change Γ for gene TP53 using the 1D model with $L = 20$. Data are divided into transition and transversions. We give standard errors of the mean (SEM) and standard deviations (σ) for each distribution. From these we estimate the probability of each distribution being a random sample from the set of all mutations, $p_{\rm all}$, or being a sample from a population similar to the pathogenic mutations, $p_{\rm pa}$ (cp. Fig. S7). There are 224 silent transitions and 142 silent transversions; 67 intronic transitions and 46 intronic transversions. The pathogenic mutations and all possible mutations outnumber the silent and intronic populations by factors of 10–1000 and so it is the SEM for the smaller populations that is significant. It is clear that the mean CT change $\overline{\log_{10}\Gamma}$ for the silent and intronic populations is far more similar to the pathogenic populations than to the entire population of all possible mutations. This is true for both transitions and transversions, although the p-value for the intronic transitions is not statistically significant (i.e. ≥ 0.05) which we attribute to the small number of available intronic data.

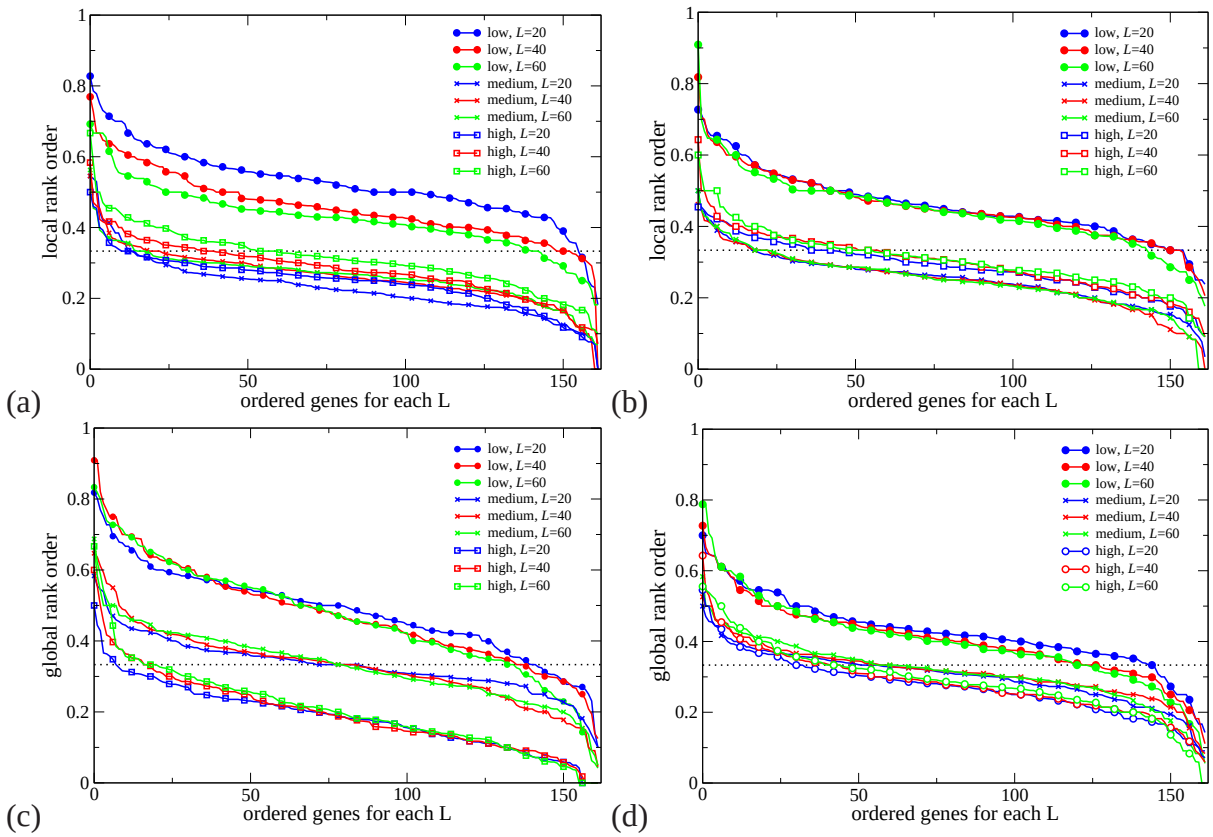

Figure 1: Sorted prevalence of the low, medium and high CT change among *local* (a+b) and *global* (c+d) rankings for pathogenic mutations in 162 genes using the 1D (a+c) and the 2-leg (b+d) models. Results are consistent for all three lengths $L = 20, 40, 60$. The $1/3$ value expected by chance is shown as a dashed horizontal line. Low rankings are dramatically more prevalent locally and globally than chance would suggest.

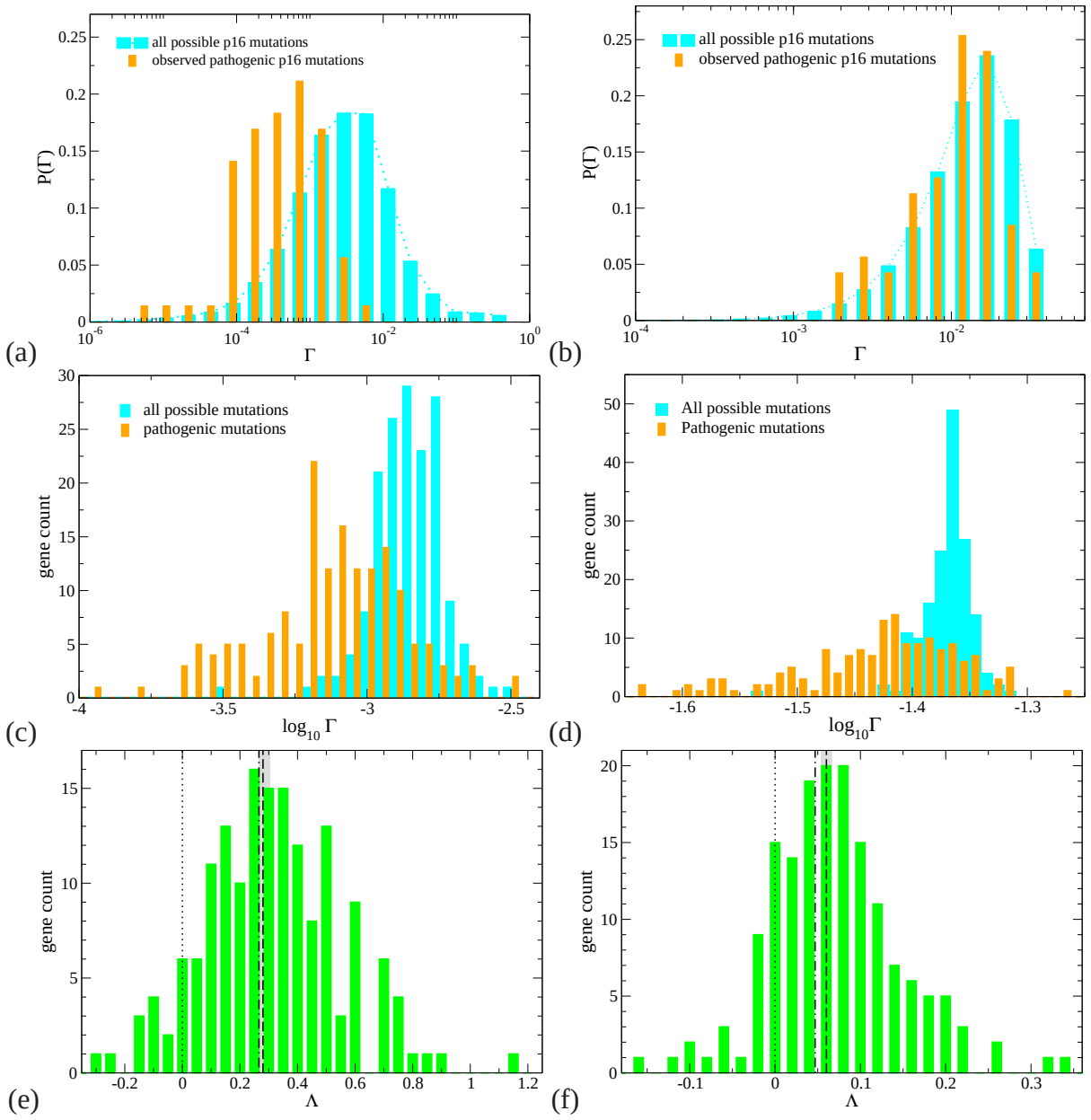

Figure 2: (a+b) Distribution of the change in charge transport Γ for pathogenic (orange bars) and all possible (cyan bars) mutations for the $p16$ (CDKN2A) gene with 26740 base pairs and 111 known pathogenic mutations. (c+d): Distribution of the average (logarithmic) change in charge transport $\langle \log \Gamma \rangle$ for all pathogenic (orange bars) and all possible (cyan bars) mutations for all 162 genes. (e+f): Distribution of the global shift Λ values for all genes, showing a consistent tendency to positive values. The average $\bar{\Lambda}$ (dashed) and weighted average $\tilde{\Lambda}$ (dash-dotted) values are indicated by vertical lines similarly to the 0 line (dotted). The grey bars denote the error of mean for $\left\langle \bar{\Lambda} \right\rangle$. The results for the 1D and 2-leg models are displayed in panels (a,c,e) and (b,d,f),

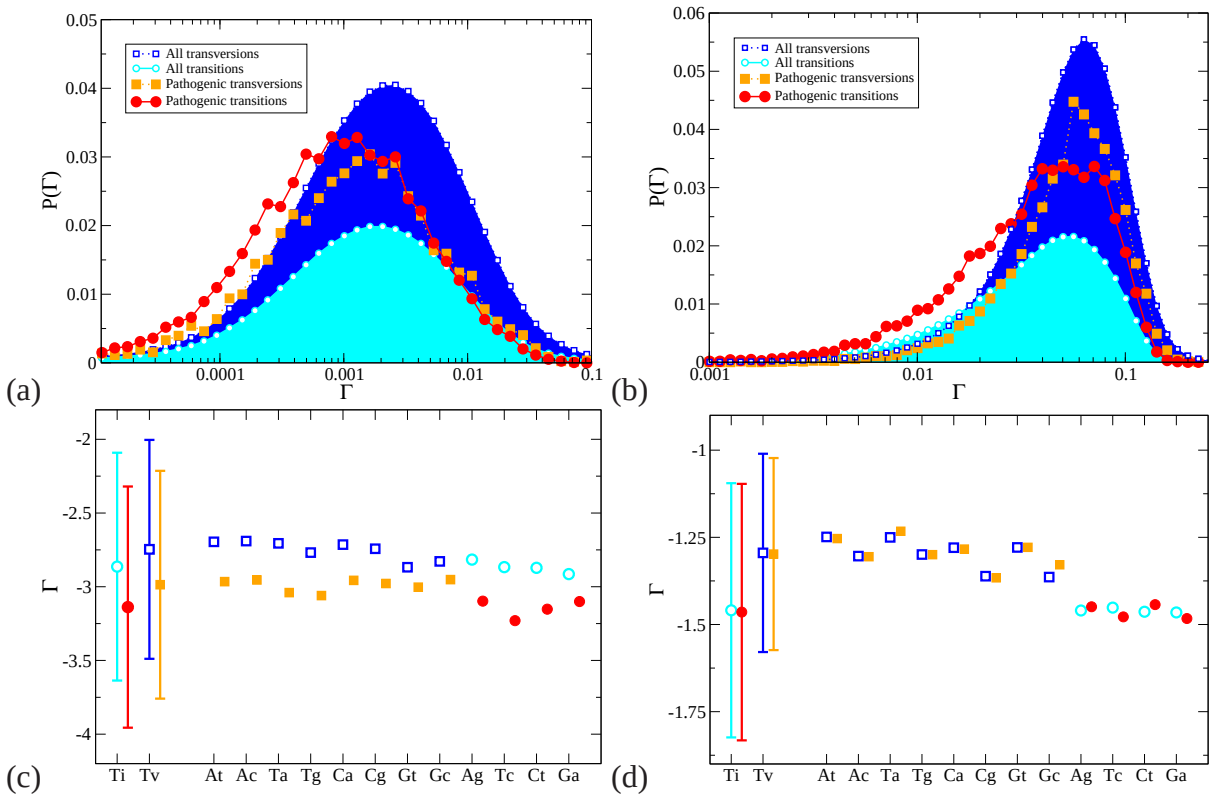

Figure 3: Distributions of Γ for the 1D (a) and 2-leg (b) models for all genes, with mutations divided into transitions and transversions. The distributions are normalised by the size of the mutation dataset. Lines are guides to the eye only. The means (symbols) and standard deviations (error bars) of the distributions of $\log \Gamma$ are shown in panels (c) and (d) for the 1D and 2-leg models. *Estimated errors of the means are smaller than the symbols.* Distributions are shown for transition (Ti) and transversion (Tv) mutations, and for the twelve types of point mutation individually. Open symbols (blue, cyan) are for the set of all mutations, filled symbols (orange, red) for the set of pathogenic mutations.

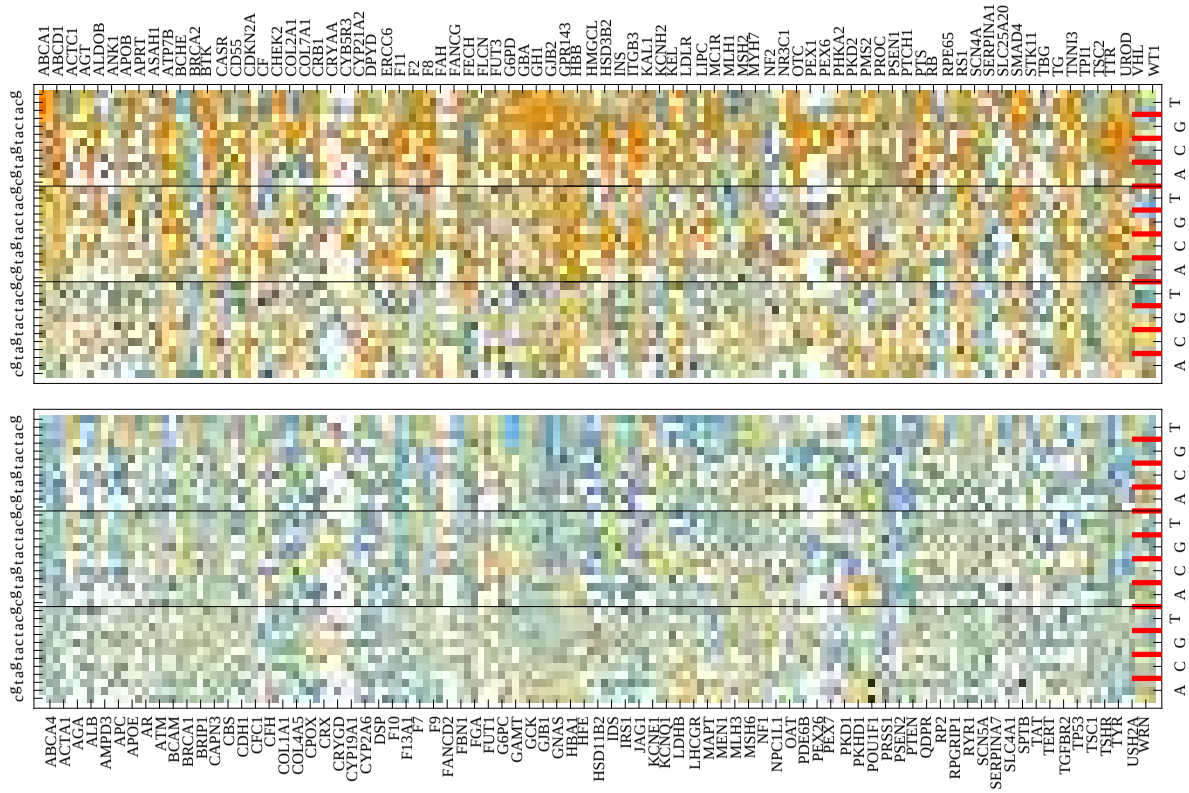

Figure 4: Distribution of subset shifts λ for the 1D (top) and 2L (bottom) models over all 162 genes split into the 12 possible mutations (Ac, Ag, At, Ca, ..., Tc, Tg). The capital letters on the right axes denote the original base pairs, whereas the lowercase letters in the left axes show the mutant base. The short red tick marks on the right axes distinguish different original bases. The system sizes $L = 20$, 40 and 60 are shown in the bottom, centre and top row for each model. The orange shading corresponds to positive λ and blue to negative. The white squares correspond to cases for which either no corresponding pathogenic mutations are known (1029 cases) or for which the subset shift is inconclusive (3 cases for the 2-leg model).

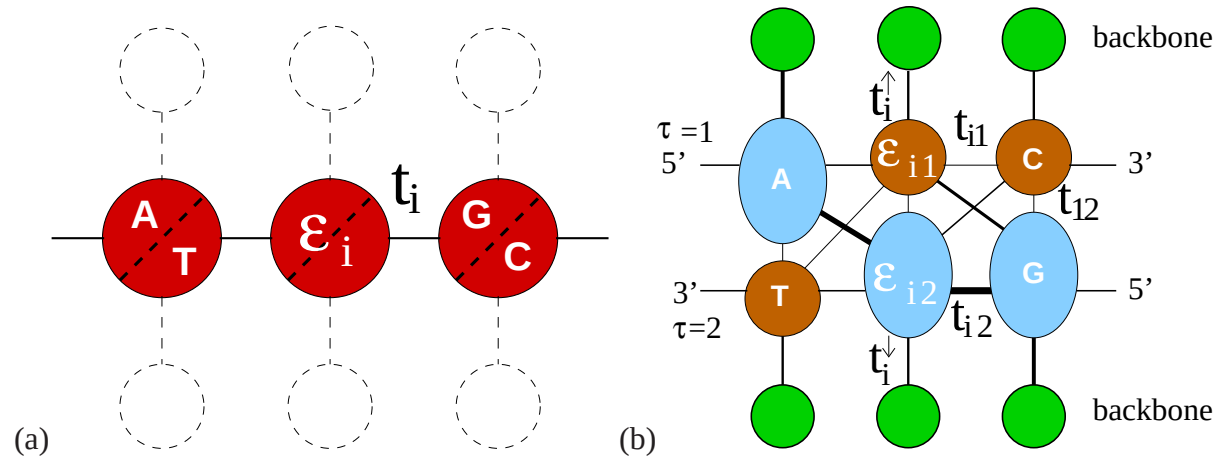

Figure 5: Schematic models for charge transport in DNA. The nucleobases are given as circles (red, denoting pairs) and ellipses (blue, brown for single nucleotides). Electronic pathways are shown as solid lines of varying thickness to indicate variation in strength. Model (a) indicates the 1D model where the sugar-phosphate backbone is ignored. In model (b), brown circles denote the smaller pyrimidines, blue ellipses are the large purines and green circles denote the sugar-phosphate backbone sites. Note that diagonal hopping between purines is favoured, and between pyrimidines disfavoured, by the larger size of the purines.

Supplementary Material

Comparing the Averaged Electronic Properties for the Pathogenic and Non-pathogenic Mutations for Each Gene

We denote the genomic sequence of a gene with length $\mathcal{N}$ base pairs (bps) as $(s_1, s_2, \cdots, s_{\mathcal{N}})$. Each point mutation of a given gene is characterized by the set (k, s), where k and s are the position of the point mutation in the genomic sequence and the mutant nucleotide which replaces the nucleotide s_k of normal DNA, respectively. There are totally $3\mathcal{N}$ possible point mutations of a gene with $\mathcal{N}$ bps. The sets of these $3\mathcal{N}$ mutations and the pathogenic mutations for the gene are denoted as $M_{\rm all}$ and $M_{\rm pa}$, respectively. $M_{\rm pa}$ is a subset of $M_{\rm all}$. For every possible point mutation, we compute the *mean* quantum mechanical transmission coefficient $T_L^{(k)}$ of a subsequence with length L of the *wild-type* gene. Here the mean is determined by averaging over all individual transmission coefficients $T_{j,L}$ with $j = k - L + 1, k - L + 2, \ldots, k$. In this way, the influence of the full neighborhood of hotspot k is taken into account and not just the mutation itself. The results of $T_L^{(k)}$ for $k \in M_{\rm pa}$ already show some signatures of atypical CT reponse for the 1D model.[38] However, the signal is much less pronounced in the 2-leg model. Hence we study the *difference* in CT between a healthy DNA base and the 3 possible mutations. For example the hotspot 14585 of $p53$ contains the correct C/G base pair in the wild but of the three possible mutations $C/G \to G/C$, $C/G \to A/T$ and $C/G \to T/A$ only the last one is know to lead to cancer.[5] Averaging again over all incident energies and subsequences of length L containing the

hotspot (k, s), we can characterize the *average change* in CT as

$$\Gamma_{L,q}^{(k,s)} = \frac{1}{L} \sum_{j=k-L+1}^{k} \int_{E_0}^{E_1} \frac{|T_{j,L}(E) - T_{j,L}^{(k,s)}(E)|^q}{E_1 - E_0} dE \quad . \tag{3}$$

with $q = 1$ or 2. We find that results for $q = 1$ and 2 are similar. Hence in the manuscript we restrict our discussion to $q = 2$. We calculate such Γ estimates for all possible $3\mathcal{N}$ mutations of each gene and compare the probability distribution of CT change $\Gamma_{L,q}^{(k,s)}$ for $(k, s) \in M_{\rm all}$ and $(k, s) \in M_{\rm pa}$ for each gene. The result for the $p16$ gene was shown in Fig. 2(a) as an example. As a control group, we also shuffled the $p16$ sequence randomly under the conditions that (1) the contents of the 4 bases are not changed, and (2) the positions of the mutations can be moved but the numbers of the 12 types of mutations are not changed. The distributions of the averaged Γ for 1D and 2-leg models with $L = 40$ of the 20 shuffled sequences are shown in Fig. S1. It is clear that the distributions of Γ for the M_{all} and M_{pa} are almost identical.

CT Change for the 12 Type of Mutations

The comparison of Γ between the pathogenic and all possible mutations for the 12 types of point mutations is shown in Fig. S2. It is clear for the 1D model (a–l) Γ tends to be smaller for the pathogenic mutations. However, the difference is not visible for the 2L model (m–x).

Local ranking of point mutations at hotspot sites

In order to study the local effects of pathogenic mutations on CT, we compare $\Gamma_{L,2}^{(k,s)}$ of each pathogenic mutation (k, s) with the other two non-pathogenic ones at the same position k and de-

termine the *local ranking* (LR) of CT change for (k, s). There are three possibilities of LR, namely *low*, *medium* and *high*. Note that those hotspots k with more than one pathogenic mutations are excluded in the LR analysis. As an example, percentages of the three LR for the pathogenic mutations of $p16$ are shown in the left panels of Fig. S3. of pathogenic mutations with low CT change are evidently larger than the medium and high ones for all L. Let us again ask how significant this tendency is across all 162 genes. Figure S4 shows similar ranking analysis results as in Fig. S3 but now for *all* $M_{\rm pa}$. We see that the tendency towards low CT change in the pathogenic mutations is quite strong overall. In Fig. 1 we have sorted the LR ranking for each gene according to prevalence. We find that for $L = 20$, 40 and 60 the low CT change corresponds to 155 (95%), 148 (91%) and 140 (86%) of all 162 genes with pathogenic mutations. Note that similarly consistent is the result for large CT with only about 30 of all genes having high CT change.

Global CT rankings at hotspot sites

Another way to compare the CT change is a *global* ranking (GR). We have sorted the CT change $\Gamma_{L,2}^{(k,s)}$ for *all* possible $3\mathcal{N}$ mutations of a gene with $\mathcal{N}$ bps in order to get a ranking of *every* pathogenic mutation (k, s). By dividing each ranking by $3\mathcal{N}$ we compute the normalised GR $\gamma_{L,2}^{(k,s)}$ of the mutation with values between 0 and 1. As before for $\Gamma_{L,q}^{(k,s)}$, smaller values of $\gamma_{L,q}^{(k,s)}$ mean smaller CT change. To characterise the CT change in a quantitative way, we divide the $\gamma_{L,2}^{(k,s)}$ of the pathogenic mutations into again three groups as before, i.e. low ($\gamma < 33.3\%$), medium ($33.3\% \leq \gamma < 66.7\%$), and high ($\gamma \geq 66.7\%$) CT change. The distributions of the GR for the complete set of pathogenic mutations of $p16$ is shown in Fig. S3 as an example. As for the

LR results, the pathogenic genes lead to many $\gamma_{L,2}^{(k,s)}$ values with low CT change. This is most pronounced in the 1D model as shown in Fig. S3(c). The results of the GR for the 162 genes are shown in the bottom row (c) and (d) of Figs. S4 and 1. We see that the GR results are fully consistent with the LR rankings.

Consistency of CT rankings for all DNA sequences

The prevalence ordering as shown in Fig. 1 does not imply that the order of the genes themselves is the same in all parts (a), (b), (c) and (d) of the figure. Therefore we have calculated the correlations in the ordering and found that in both models and across models and for all $L = 20$, 40 and 60, we find positive correlation coefficients. Hence genes which have a low change in CT for, e.g., the local ranking at $L = 20$, also retain this low rank for the other L values as well as the global ranking. Similarly, this positive correlations implies that in those few case where the mutations in a gene lead to high CT change, they do so across all local as well as global rankings. This confirms that our results are internally consistent.

We graphically summarise the results for all 162 disease-related genes in Fig. S5. For each gene, we have shown a positive deviation from the 0.33 line by orange —supporting the scenario of small CT change for pathogenic mutations — and by blue when the results seem to show no or negative indication with CT change. The criteria corresponds to local and global ranking results for $L = 20$, 40 and 60 for the 1D and the 2-leg models. Similarly, in Fig. S6, we average of all 12 criteria and show the resulting, overall agreement with the CT hypothesis: 161 of 162 genes are

above the 33% line and hence show that for both 1D and 2-leg model and averaged over lengths 20, 40 and 60, a small CT change correlates with the existence and position of pathogenic mutations. Only for STK11 do we see that there is no overall agreement.

Difference and similarities in the two models

The 2-leg model[16] allows inter-strand coupling between the purine bases in successive base pairs, in accordance with electronic structure calculations,[36] and should therefore be a better model for bulk charge transport along the DNA double helix; the 1D model, by contrast, makes use of the site energies of only the bases on the coding strand,[15] and so is most representative of the electronic environment along that strand. We also find that the 2-leg model recovers some of the coding strand dependence of the 1D model upon decreasing the diagonal hoppings. For 28 genes, we find that reducing only the diagonal hopping elements by $1/2$ leads to a much greater agreement with the 1D results similar to Fig. 3(c).

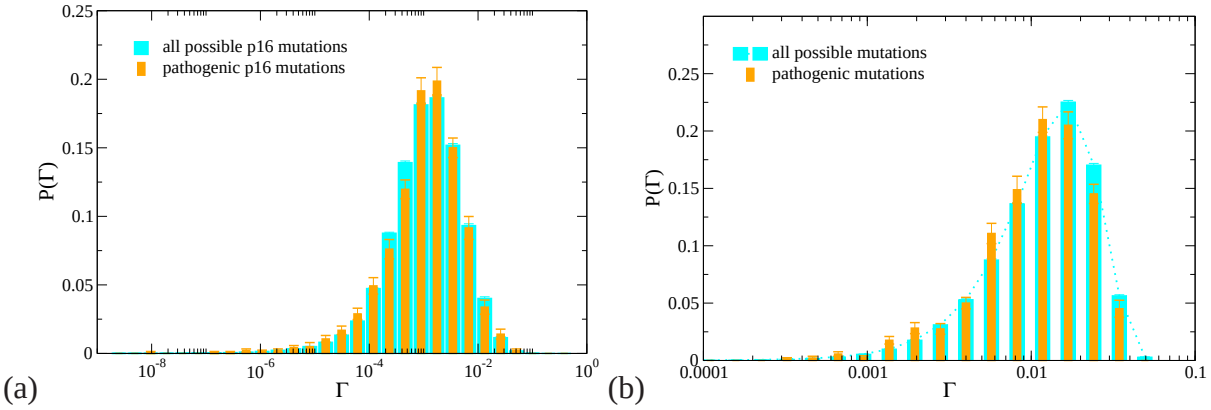

Figure S1: (Supplementary) Distribution of the change in charge transport in (a) 1D and (b) 2L models Γ for pathogenic (orange bars) and all possible (cyan bars) mutations averaged for the 20 shuffled $p16$ (CDKN2A) DNA strands with 26740 base pairs. All results shown are for $L = 40$, data for $L = 20$ and 60 are similar.

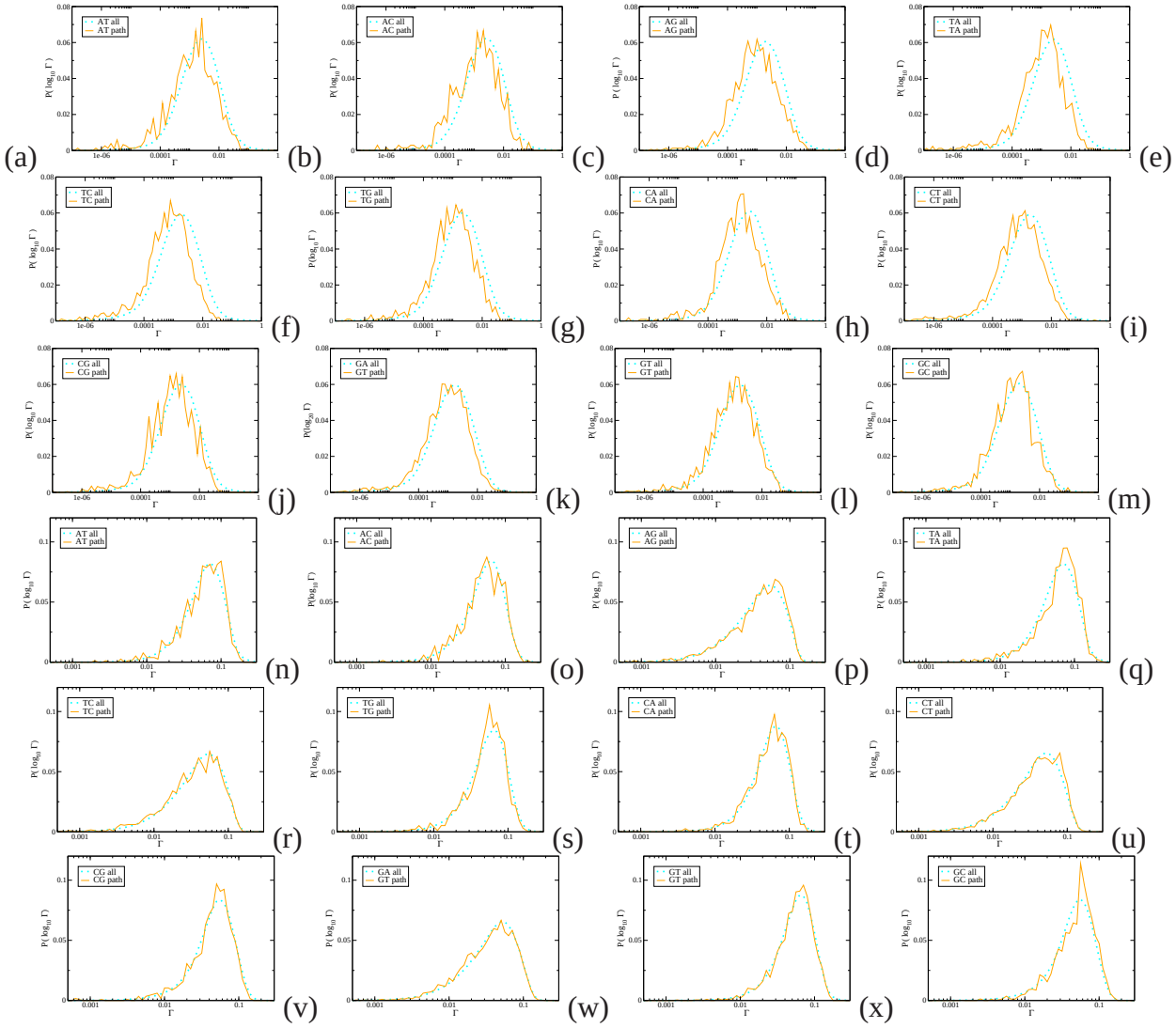

Figure S2: (Supplementary) Panels a-l: 1D model, results divided into the twelve subtypes of mutation. The shift for pathogenic mutations is clearly present in every case. Panels m-x: 2L model, results divided into the twelve subtypes of mutation. There is no consistent shift for pathogenic mutations.

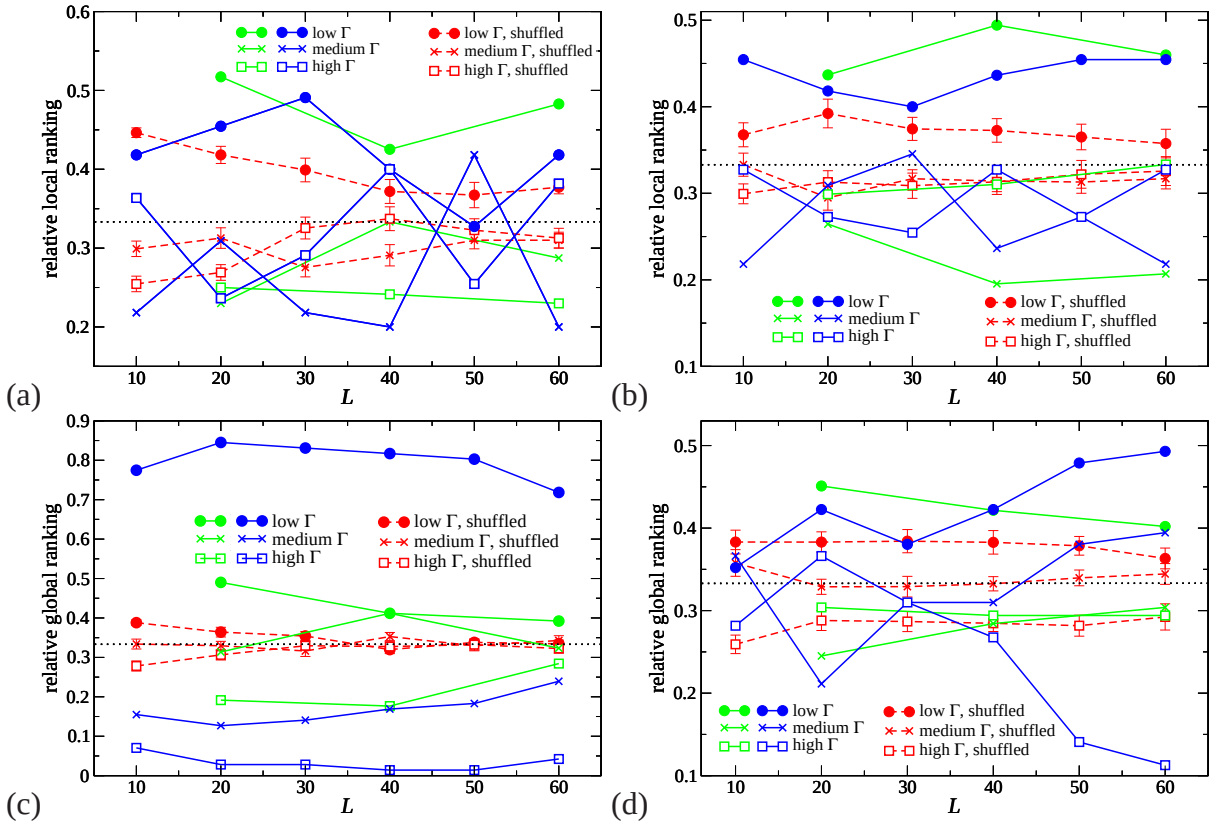

Figure S3: (Supplementary) Distribution of the *local* (a+b) and *global* (c+d) ranking results of pathogenic mutations of $p16$ (CDKN2A) (blue solid lines) and CYP21A2 (green) as a function of window lengths L. The dashed lines indicate averaged results for 20 randomly shuffled $p16$ sequences. The left/right columns distinguish results for the 1D/2-leg models. The dashed horizontal line shows the 33% mark expected for a completely random sequence. All lines are guides to the eyes only. Error bars are within symbol size.

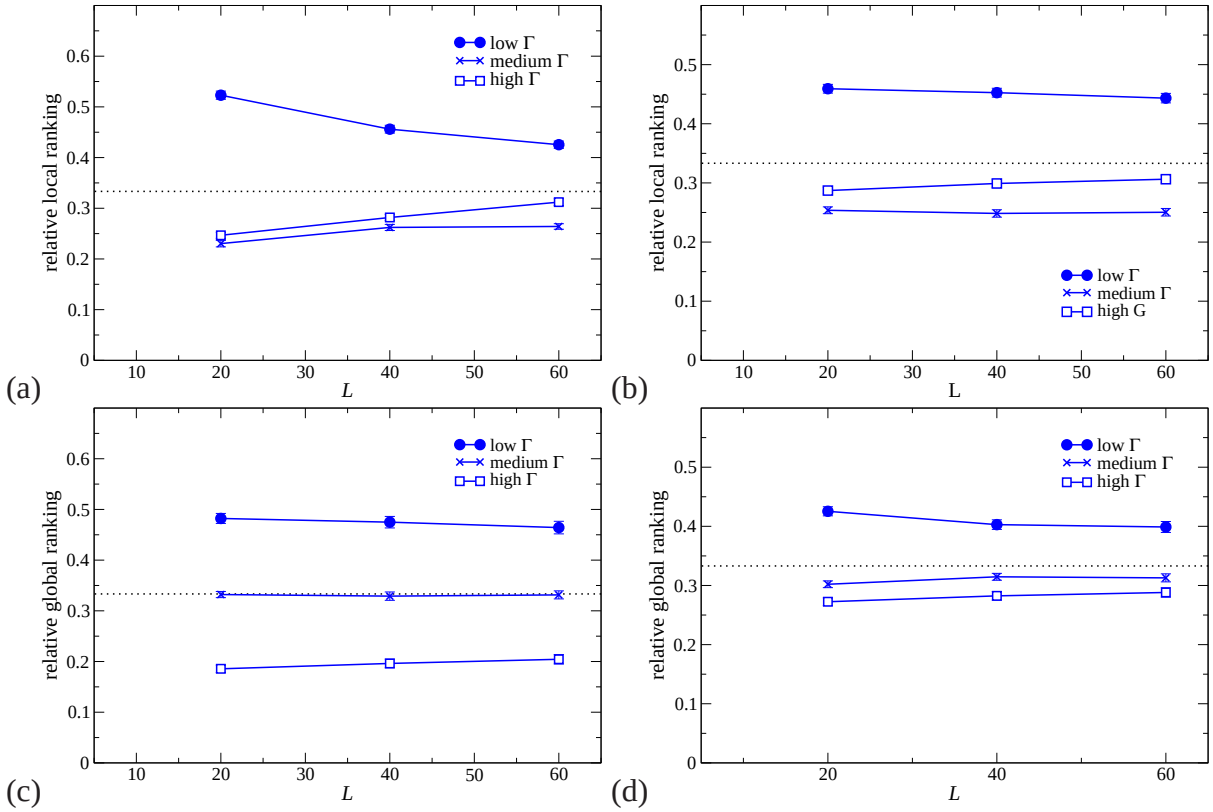

Figure S4: (Supplementary) Distribution of the *local* (a+b) and *global* (c+d) ranking results of *all* 19882 pathogenic mutations of the 162 genes as a function of window lengths L. The left/right columns distinguish results for the 1D/2-leg models. The dashed horizontal lines show the 33% mark of a completely random sequence. All lines are guides to the eyes only.

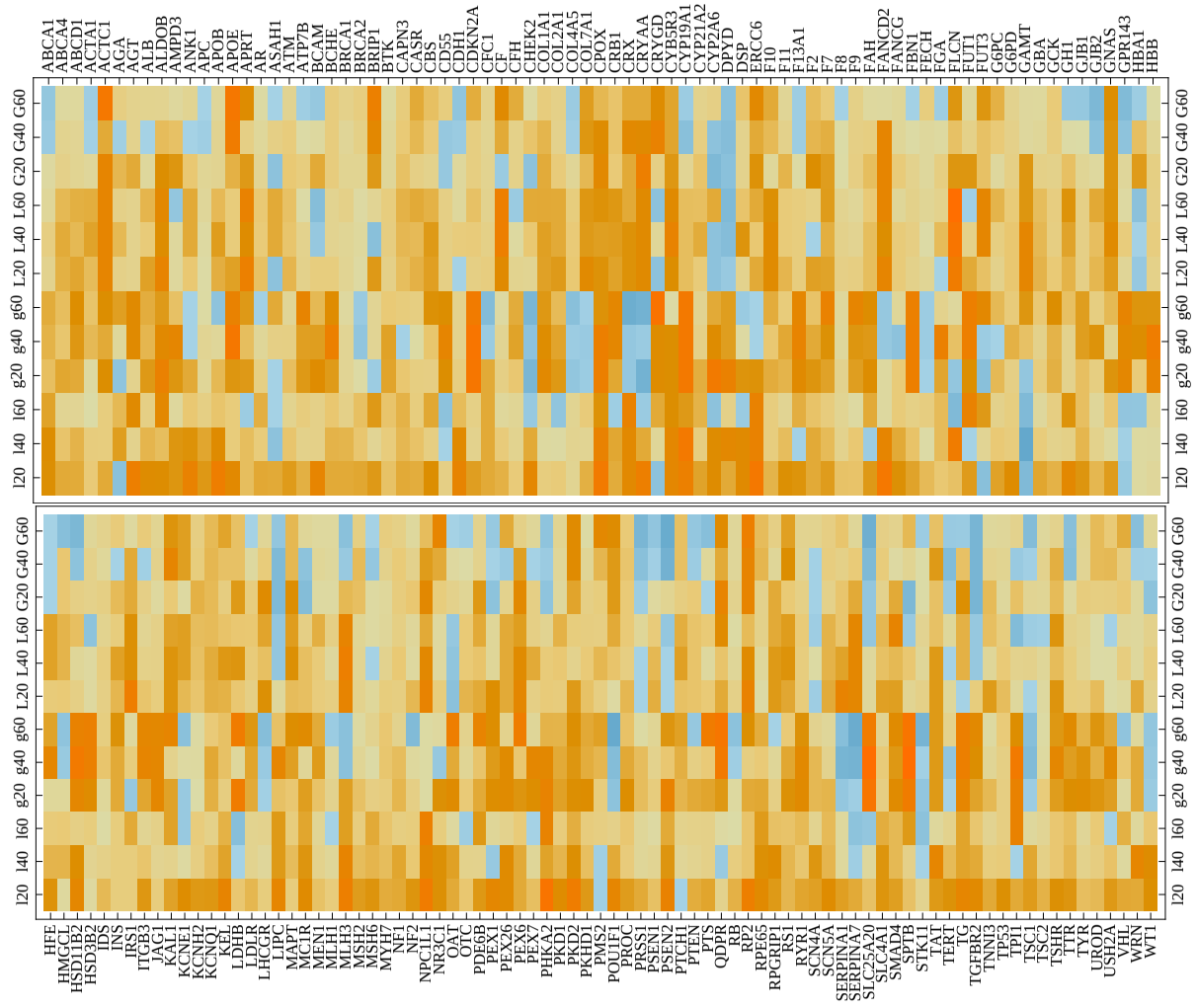

Figure S5: (Supplementary) Numerical representation of the 12 criteria for all 162 genes, i.e. deviation from the 0.33 line for the *local* rankings (li, Li) and the *global* rankings (gi, Gi) corresponding to the sorted prevalence for $L = 20, 40$ and 60, respectively. The lower case (l,g) indicates results for the 1D model, uppercase (L,G) refers to the 2-leg model. The genes are named according to the usage in the DNA databases.[3–6] The orange shading corresponds to an agreement with the CT hypothesis while the blue shading denotes disagreement. The first (last) column in the top (bottom) row gives the scale from 0 to 1 with 0.33 corresponding to the white square.

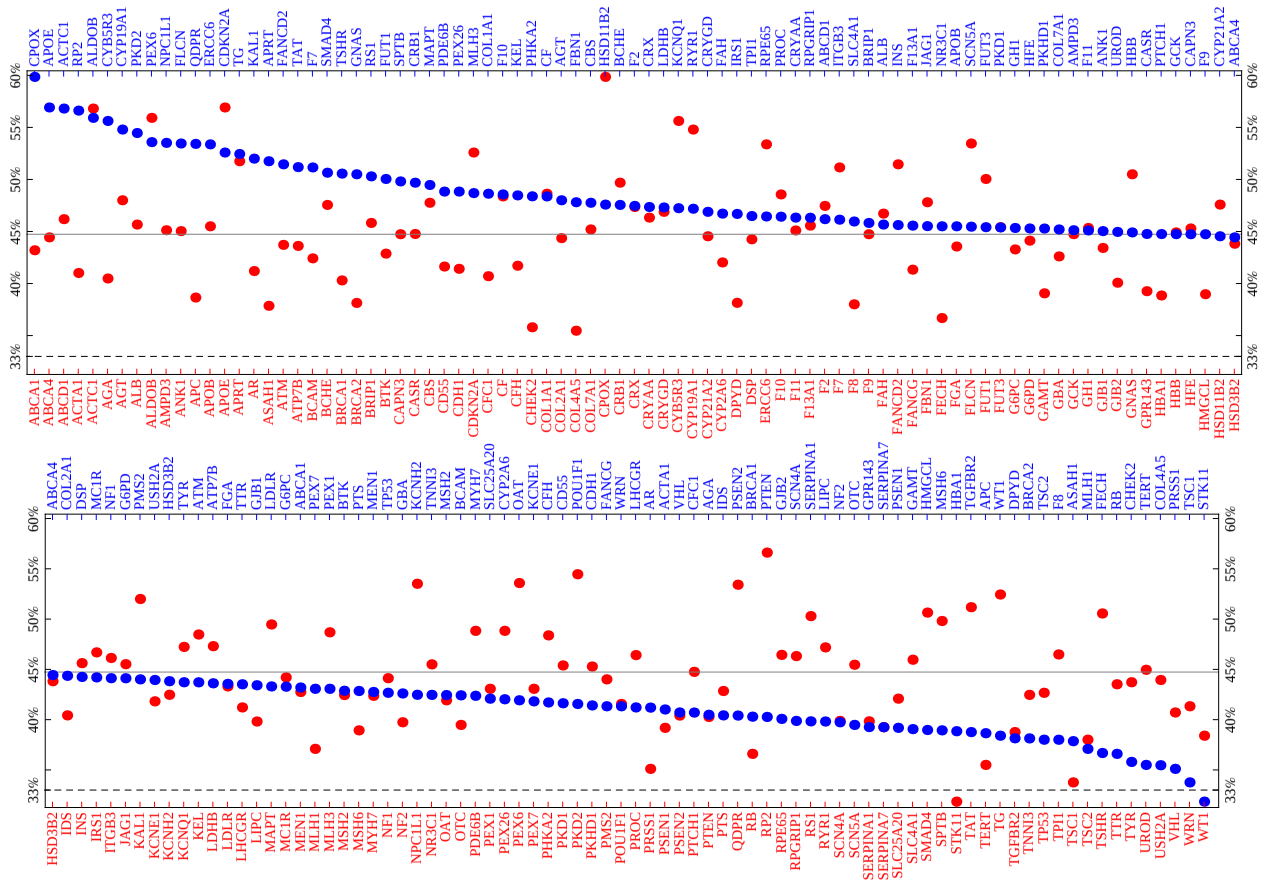

Figure S6: (Supplementary) Graphs of the *average* over all 12 criteria as displayed in Fig. S5. The red data points and gene names correspond to an alphabetic ordering of genes, whereas the blue points and labels are ordered according to the magnitude of the average. A larger average denotes a better agreement with our hypothesis. Points which lie below the dashed 33% line show genes which on average fail. The average over all genes is denoted by the solid line. Results for HSD3B2 (unsorted) and ABCA4 (sorted) have been duplicated in both rows.

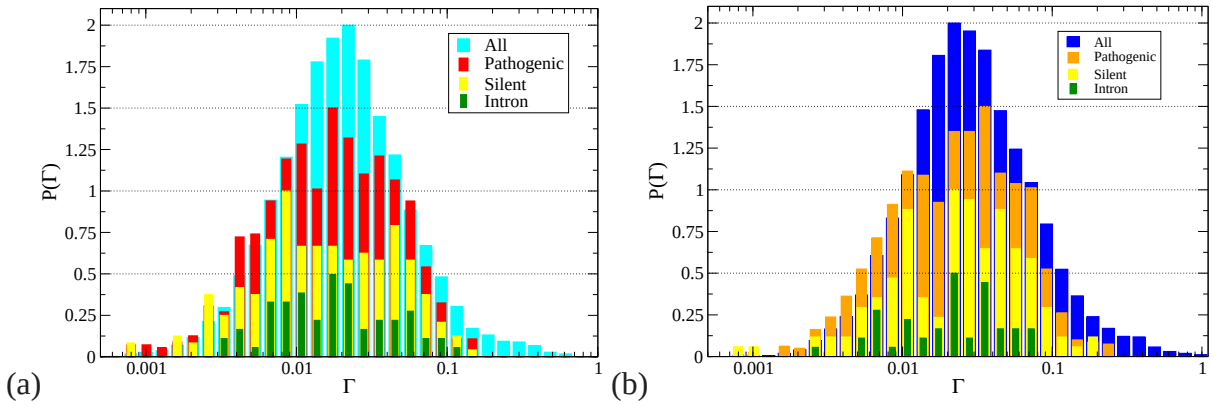

Figure S7: Histograms of Γ distributions for (a) transitions and (b) transversions in TP53, simulated using the 1D model and $L = 20$. Histograms are shown for all possible mutations and for pathogenic, silent and intronic subsets. The maximum heights of the populations are scaled to be 2, 1.5, 1 and 0.5 to ease comparison. The scales factors are indicated by the dotted horizontal lines.

Table S1: (Supplementary) List of the 162 genes with their lengths (bps), number of all point mutations (N_{pa}), and their numbers of the 12 types of point mutations. For example, N_{At} means the number of $A \to T$ substitution.

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
ABCA1	147154	87	0	4	9	2	7	2	4	24	3	18	7	7
ABCA4	128313	382	11	9	21	13	51	21	27	73	19	99	23	15
ABCD1	19894	223	8	7	14	6	31	3	15	46	17	47	13	16
ACTA1	2852	164	10	7	22	5	13	6	13	12	11	29	17	19
ACTC1	7631	14	0	1	3	0	0	0	0	4	1	2	1	2
AGA	11668	19	0	0	0	1	3	1	0	2	0	8	2	2
AGT	11673	10	0	0	1	0	1	1	0	5	0	1	0	1
ALB	17127	63	3	2	13	2	1	0	1	6	1	24	4	6
ALDOB	14448	28	0	0	0	1	9	1	3	5	3	3	1	2
AMPD3	56903	11	0	1	0	1	1	0	0	6	1	0	0	1
ANK1	144397	18	0	0	1	0	2	0	1	7	1	4	2	0
APC	108353	222	10	0	4	18	1	8	21	83	28	18	28	3
APOB	42645	51	0	0	2	4	1	1	3	26	2	8	3	1
APOE	3612	33	0	1	1	0	2	0	2	9	2	9	2	5
APRT	2466	13	2	0	1	0	3	0	0	1	0	4	1	1
AR	180246	299	11	6	24	11	31	12	22	53	25	56	31	17

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
ASAH1	28574	12	1	0	3	1	0	0	1	0	3	1	0	2
ATM	146268	169	8	3	20	9	11	15	5	55	10	19	8	6
ATP7B	78826	315	10	14	25	14	27	10	17	62	16	68	30	22
BCAM	12341	14	1	0	1	1	0	0	1	4	1	5	0	0
BCHE	64562	58	6	2	6	3	6	3	2	12	0	8	5	5
BRCA1	81155	301	12	6	30	14	29	23	12	63	15	38	50	9
BRCA2	84193	162	12	9	20	8	11	8	12	33	13	15	19	2
BRIP1	180771	13	1	0	0	0	1	1	0	3	2	2	1	2
BTK	36741	329	15	14	29	19	47	23	26	44	14	48	32	18
CAPN3	64215	213	2	9	18	5	23	6	10	45	19	48	14	14
CASR	102813	144	2	5	12	4	21	7	8	20	10	38	12	5
CBS	23121	107	2	1	6	4	10	0	4	24	7	39	2	8
CD55	38983	14	0	0	1	2	0	1	0	4	0	3	1	2
CDH1	98250	30	0	1	2	0	2	2	0	9	1	8	4	1
CDKN2A	26740	71	1	3	4	2	6	6	5	12	3	11	8	10
CFC1	6748	10	0	0	0	0	1	0	0	4	1	4	0	0
CF	188699	828	35	31	103	50	85	54	47	117	41	136	84	45
CFH	95494	83	3	3	8	6	10	5	2	10	6	14	13	3
CHEK2	54092	20	1	1	2	0	1	0	2	4	0	7	1	1
COL1A1	17544	292	0	2	2	0	1	2	1	21	4	134	79	46
COL2A1	31538	124	0	1	2	1	1	1	5	26	0	53	19	15

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
COL4A5	257622	244	2	0	4	2	2	5	4	20	1	117	55	32
COL7A1	31088	265	0	3	6	2	1	0	1	56	7	122	34	33
CPOX	14152	36	0	2	1	0	3	1	0	14	2	9	3	1
CRB1	210178	91	3	1	2	8	16	7	3	11	2	22	11	5
CRX	21483	18	0	1	1	0	0	1	2	4	0	8	0	1
CRYAA	3773	10	0	0	0	0	0	0	1	5	0	3	1	0
CRYGD	2882	12	0	1	0	0	0	0	4	3	1	2	1	0
CYB5R3	30587	35	0	0	3	0	6	2	2	12	0	10	0	0
CYP19A1	129126	13	0	0	0	0	2	1	0	5	0	5	0	0
CYP21A2	3338	102	4	4	5	7	8	4	6	23	2	25	4	10
CYP2A6	6897	12	1	0	1	1	2	0	0	2	0	2	2	1
DPYD	843317	34	2	3	7	2	0	1	2	7	0	5	4	1
DSP	45077	20	0	0	2	1	1	1	0	6	1	6	1	1
ERCC6	80364	18	1	0	2	1	1	1	0	10	1	1	0	0
F10	26731	81	1	4	5	1	6	2	4	11	3	33	5	6
F11	23718	131	2	5	6	3	17	3	9	28	2	29	13	14
F13A1	176614	55	1	0	2	0	6	4	4	12	3	14	8	1
F2	20301	42	0	3	3	0	1	1	0	11	1	17	3	2
F7	14891	164	4	1	13	1	17	4	9	30	6	55	13	11
F8	186936	1168	79	47	124	56	117	78	55	153	72	198	112	77
F9	32723	707	31	26	55	58	69	52	42	54	28	135	95	62

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
FAH	33342	26	2	1	2	0	1	3	2	6	0	5	4	0
FANCD2	75502	14	0	0	0	0	3	3	0	4	0	4	0	0
FANCG	6179	16	0	0	0	0	2	1	0	7	0	2	2	2
FBN1	237414	640	18	12	52	32	88	37	21	63	32	173	68	44
FECH	38454	49	2	1	2	3	7	3	1	11	1	11	4	3
FGA	7618	45	3	1	3	3	1	2	3	12	2	7	7	1
FLCN	24971	11	0	0	1	0	0	0	0	4	2	3	1	0
FUT1	7380	22	0	0	1	2	2	1	2	5	1	4	1	3
FUT3	8587	11	0	0	0	1	0	2	2	0	0	5	0	1
G6PC	12572	66	2	2	3	2	8	3	3	13	2	15	5	8
G6PD	16182	163	3	3	21	4	15	4	8	27	15	39	13	11
GAMT	4465	11	0	2	0	0	1	0	0	1	1	3	1	2
GBA	10246	259	8	11	25	8	32	19	14	42	10	53	19	18
GCK	45153	255	5	13	15	7	32	8	19	40	11	64	23	18
GH1	1636	35	2	2	7	0	3	1	1	5	2	7	3	2
GJB1	10004	240	4	5	25	18	31	12	10	39	24	39	17	16
GJB2	5513	208	8	9	19	5	28	8	12	23	15	49	19	13
GNAS	71456	51	2	2	2	1	6	2	1	17	4	9	3	2
GPR143	40464	43	2	0	3	2	4	3	4	6	1	10	4	4
HBA1	842	73	2	5	9	2	5	2	7	6	9	8	7	11
HBB	1606	263	15	20	20	21	23	16	22	26	18	38	20	24

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
HFE	9612	27	1	2	0	0	4	1	0	3	2	7	3	4
HMGCL	23583	27	2	0	2	0	3	1	1	4	1	8	3	2
HSD11B2	6421	24	1	0	1	0	3	2	1	12	1	3	0	0
HSD3B2	7879	32	0	1	1	1	2	3	3	8	3	6	2	2
IDS	26493	203	15	8	15	2	16	13	17	31	19	32	20	15
INS	1431	30	0	0	2	0	3	2	1	3	6	6	4	3
IRS1	64538	14	0	1	3	0	1	0	1	2	1	3	0	2
ITGB3	58870	53	2	2	3	1	10	4	1	12	1	11	5	1
JAG1	36257	131	2	0	3	6	11	6	11	30	12	28	16	6
KAL1	203313	25	0	0	1	2	1	1	1	9	2	6	1	1
KCNE1	65586	17	0	0	1	0	2	0	1	5	0	6	1	1
KCNH2	32966	266	8	11	27	5	19	12	15	61	9	43	35	21
KCNQ1	404120	226	3	2	19	8	24	5	12	44	13	61	11	24
KEL	21303	33	2	0	3	1	3	0	0	9	1	13	0	1
LDHB	22501	11	1	1	1	0	1	2	1	1	0	2	0	1
LDLR	44450	741	23	31	48	31	84	35	51	88	48	168	92	42
LHCGR	68951	37	2	3	3	3	7	3	2	7	1	3	2	1
LIPC	136898	11	0	1	2	0	0	1	0	2	0	4	0	1
MAPT	133924	36	3	2	2	0	3	2	2	6	1	9	5	1
MC1R	2360	24	0	1	1	0	4	0	3	8	0	5	1	1
MEN1	7779	239	10	7	8	9	26	11	19	44	14	38	33	20

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
MLH1	57359	275	16	15	26	18	19	17	18	42	20	36	28	20
MLH3	37769	17	0	1	5	0	1	0	0	2	1	4	2	1
MSH2	80098	238	16	11	25	8	9	14	11	62	14	30	25	13
MSH6	23872	54	3	1	5	2	3	0	3	17	6	7	4	3
MYH7	22924	268	8	10	20	4	19	8	16	47	17	80	16	23
NF1	282701	338	22	4	24	20	35	26	14	82	24	44	29	14
NF2	95023	72	5	2	5	2	6	1	2	25	4	7	11	2
NPC1L1	28781	26	0	0	3	2	0	0	0	11	1	8	1	0
NR3C1	157582	14	1	0	1	1	4	1	0	1	0	4	0	1
OAT	21580	42	0	0	2	2	4	0	3	9	2	11	5	4
OTC	68968	276	16	11	28	9	31	18	17	36	15	44	27	24
PDE6B	45199	20	1	0	0	3	3	1	2	5	1	4	0	0
PEX1	41509	24	0	0	0	0	4	1	2	7	3	6	0	1
PEX26	11503	10	0	0	0	0	3	0	0	3	2	2	0	0
PEX6	15143	18	0	1	0	0	3	1	1	7	0	5	0	0
PEX7	91337	24	1	2	2	1	1	3	2	6	1	4	1	0
PHKA2	91305	23	0	1	2	0	0	1	1	11	0	5	2	0
PKD1	47189	149	2	3	6	5	12	4	8	59	10	27	8	5
PKD2	70110	35	1	0	1	1	1	1	2	17	0	7	3	1
PKHD1	472279	213	8	10	22	7	29	9	7	50	7	38	17	9
PMS2	35868	21	3	1	1	1	0	0	0	6	0	5	4	0

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
POU1F1	16954	22	1	0	2	1	3	1	1	6	0	4	2	1
PROC	10802	203	6	6	10	3	21	6	15	40	8	55	13	20
PRSS1	3592	26	1	2	2	2	2	0	2	5	1	5	2	2
PSEN1	83931	154	6	8	13	8	22	11	7	21	12	19	16	11
PSEN2	25532	18	2	2	3	0	0	0	0	5	1	5	0	0
PTCH1	73984	59	3	2	1	2	4	2	7	15	2	11	8	2
PTEN	105338	98	2	2	10	9	13	11	5	15	8	15	6	2
PTS	7595	27	1	0	8	1	0	2	0	6	2	4	2	1
QDPR	57702	20	0	1	2	0	3	3	0	3	0	6	1	1
RB	180388	226	9	8	18	12	16	11	10	38	8	51	28	17
RP2	45418	17	0	0	1	0	1	2	0	5	2	4	2	0
RPE65	21136	42	1	1	2	1	5	3	3	9	1	7	7	2
RPGRIP1	63325	24	0	2	5	1	0	0	0	7	0	5	3	1
RS1	32422	93	3	0	7	5	11	4	5	15	5	19	7	12
RYR1	153865	244	5	4	21	9	20	6	10	56	14	63	17	19
SCN4A	34365	43	1	0	5	2	3	1	4	7	3	12	2	3
SCN5A	101611	226	0	2	18	9	16	6	13	49	10	77	15	11
SERPINA1	12332	29	4	1	0	1	2	1	2	8	1	9	0	0
SERPINA7	3870	16	1	0	0	2	1	0	1	4	0	5	1	1
SLC25A20	41966	11	0	0	1	0	0	0	0	4	1	3	1	1
SLC4A1	18428	65	1	1	3	2	5	0	6	20	4	20	1	2

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
SMAD4	49535	20	0	1	1	0	0	2	1	6	3	4	1	1
SPTB	76865	18	0	0	2	2	2	2	0	6	1	0	1	2
STK11	22637	62	4	4	2	1	4	5	7	12	5	8	8	2
TAT	10242	11	0	0	0	0	1	1	0	5	1	2	1	0
TERT	41881	30	0	1	3	1	2	0	0	10	3	8	0	2
TG	267939	33	0	1	2	1	2	1	1	7	0	14	4	0
TGFBR2	87641	14	0	0	1	1	1	0	0	5	0	3	1	2
TNNI3	5966	30	0	1	5	1	1	0	0	8	2	10	0	2
TP53	20303	2003	137	113	158	121	142	109	165	284	156	252	202	164
TPI1	3287	11	0	0	1	1	1	0	0	1	0	4	1	2
TSC1	53285	44	2	0	1	1	1	2	5	19	5	5	3	0
TSC2	40724	165	7	4	6	5	13	5	22	48	18	22	10	5
TSHR	190778	45	1	0	3	1	9	2	3	8	1	12	2	3
TTR	6944	98	4	5	10	6	15	9	6	5	1	19	11	7
TYR	117888	205	10	10	22	6	16	6	16	27	13	42	26	11
UROD	3512	45	0	1	2	5	6	2	3	9	2	11	2	2
USH2A	800503	66	0	3	1	1	2	3	6	24	3	8	10	5
VHL	10444	172	5	7	12	13	22	15	7	22	21	17	18	13
WRN	140499	22	3	1	1	1	1	0	0	11	2	1	1	0
WT1	47763	56	1	2	5	1	6	3	3	13	4	11	5	2